Muroor Vikas Rao
Project 2 Report
EECS 219

Project 2 : - MapReduce and Hadoop

In the project, first setting up of the Hadoop framework was done on the single node system. The initial stages involved following the instructions of editing the configurations of the Hadoop framework on the xml files as instructed in the discussion posted on the website. Then, running a couple of the example jar files to learn the running of the MapReduce jobs was done.

These steps were followed by developing the Java program to solve the problem stated. This involved manipulating the <key,value> pairs to obtain the output. It took for MapReduce phases to produce the file with the three sentences with the highest existence probability.

DESCRIPTION OF THE MAP-REDUCE JOBS

- The first MapReduce job involved forming the sentences and placing them into an ArrayList and adding them to the values produced as it needed to propagate through to the last stages of providing the sentence probabilities.The sentences were then tokenized and their count was calculated which was needed in the next phase to calculate the probabilities.
- The second MapReduce job involved calculating the number of sentences with at least i words, in order to calculate the value of N in the probability equation. The number of sentences with at least i words was equivalent to the number of words in the i position and this was easily calculated. In this way the probability of the words were calculated in this MapReduce phase.
- The third MapReduce phase involved calculating the probability of the sentences which was simply the product of the probabilities of the words in the sentence in the previous stage. This was easily calculated as we had the sentences carried through the previous stages in the value. This was in the form of a List.
- Next, the sentences needed to be sorted and placed in the fashion so as to display only the top3 sentences with their probabilities. This was done using TreeMap and the cleanup function which would take care of the Mappers output of the Top3 sentences. Other details have been commented in the code provided.

RUNNING THE CUSTOM JAR FILE ON THE EMR

- First for a S3 bucket was made with folders for the job (with the JAR file) , input (with the txt file).
- Then, on the EMR Management console, a new cluster was created with the configurations consisting of the path to the Custom JAR file in the S3 bucket. The input and output folder paths were also set to the s3n://pathname .
- Also, a log folder was created to ensure that the logs where stored in the S3 bucket in case of errors.
- Running the job produced the output as expected.

Note: the output folder should not already exist for the Hadoop job to run successfully.

As asked for , the script file with the given file structure was also ran using the test.sh. Certain changes were made to the script depending on the path of the hadoop on my system.
This too ran as expected.